WHEN MORE FOUNDATION MODELS MEAN LESS: DIAGNOSING AND ADDRESSING MULTI-VIEW FUSION FAILURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation-model hubs turn multi-view fusion into a selection problem: from a large heterogeneous encoder pool, which views should be fused, and how many? We show that downstream performance is non-monotonic in the number of fused encoders; later views can be redundant or task-misaligned, causing accuracy to saturate or decline. We formalise this setting as *view-set composition* and propose KAGES (Kernel-Alignment Greedy Encoder Selector), a label-aware method that orders frozen encoders by their marginal gain in centred kernel-target alignment. KAGES requires no downstream classifier training during selection, evaluates each candidate in $\mathcal{O}(n^2)$ time independent of encoder dimension, and admits a conditional $(1 - e^{-\gamma})$ prefix-wise guarantee under monotonicity and a positive submodularity ratio. Across five recognition regimes and low-shot, larger-pool, and full-data protocols, KAGES improves average AULC over full fusion by 3.9, 5.8, and 3.3 points, respectively. The same selection effect extends to image retrieval and frozen-LLM fusion. These results show that effective large-pool fusion depends on selecting a compact, task-aligned set of views rather than indiscriminately fusing more encoders.

1 INTRODUCTION

Classical multi-view learning assumes a small, predefined collection of sensors or feature views (Xu et al., 2013; Yan et al., 2021; Trosten et al., 2021). Frozen foundation models change the setting: each encoder induces a view of the same input, and model hubs supply heterogeneous pools spanning architectures, objectives, supervision, and modalities (Radford et al., 2021; Oquab et al., 2024; He et al., 2022; Liu et al., 2022). The view set therefore becomes a design variable. We call the resulting problem *view-set composition*: selecting both which encoder-induced views to fuse and how many to retain.

Existing multi-encoder systems demonstrate the benefit of fusion (Shi et al., 2025; Tong et al., 2024; Kar et al., 2024; Zong et al., 2024), but usually operate on fixed or manually specified pools. Scaling the pool exposes a missing decision: more views are not always better. Figure 1 shows that even a favourable encoder order can peak early because a later view repeats captured information or contributes task-irrelevant structure. Per-encoder transferability scores (You et al., 2021; Bao et al., 2019; Pandy et al., 2022) ignore these set interactions, while pure diversity cannot distinguish useful novelty from task misalignment. Exhaustive or validation-trained subset search is also impractical at hub scale.

KAGES addresses this gap by constructing a complete task-aware ordering of frozen encoders. At each step it adds the candidate that maximally increases centred alignment between the fused representation kernel and the label kernel. Linear-kernel additivity makes candidate evaluation independent of encoder dimension and avoids downstream retraining. The resulting objective exposes an adaptive relevance-redundancy trade-off and, under explicit monotonicity and submodularity-ratio assumptions, yields a conditional $(1 - e^{-\gamma})$ guarantee for every greedy prefix. Full derivations and proofs are in Appendix A.1.

Across five recognition regimes, KAGES front-loads useful encoders, rejects task-misaligned paradigms, and consistently improves prefix-path quality over full fusion, random ordering, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

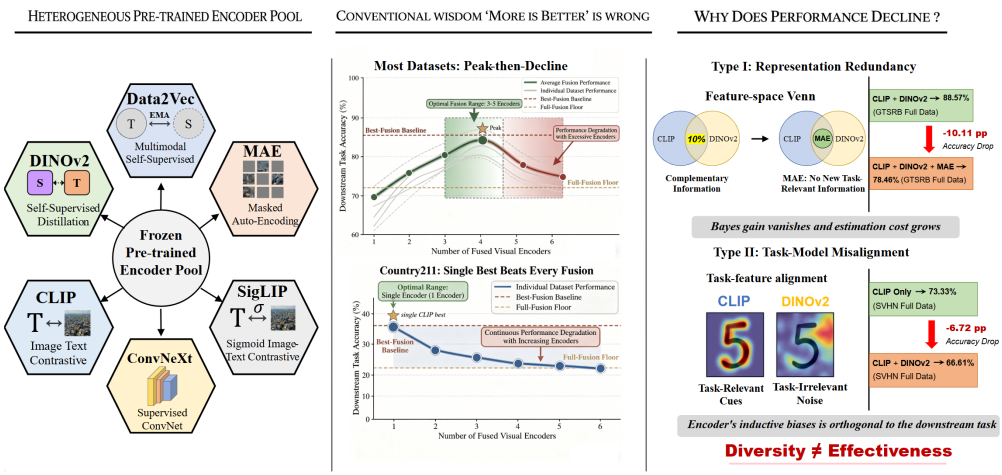

Figure 1: **View-set composition.** A heterogeneous frozen-encoder pool can exhibit peak-then-decline as views are added. Redundancy and task-model misalignment explain why diversity alone is insufficient; KAGES instead selects views by task-aligned marginal utility.

diversity selection. It also matches or exceeds stronger DPP and facility-location baselines and transfers to retrieval and frozen-LLM fusion. Our contributions are:

Contributions.

- We formulate large-pool fusion as *view-set composition* and characterise its order-specific peak through a gain-cost crossing.
- We introduce KAGES, an $\mathcal{O}(n^2)$-per-candidate set-level selector with an exact marginal-gain interpretation and a conditional prefix-wise guarantee.
- We validate the method across three classification protocols, five recognition regimes, retrieval, and frozen-LLM fusion; complete proofs, protocols, and results appear in the appendix.

2 Peak-then-Decline and the Need for View-Set Composition

We model large-pool fusion as an ordered path and explain its peak through the competition between fresh task signal and finite-sample cost.

2.1 Ordered Marginal Utility

Let $\mathcal{M} = \{m_1, \ldots, m_M\}$ be frozen encoders and π an ordering. Its prefixes

$$S_k^\pi = \{m_{\pi_1}, \ldots, m_{\pi_k}\}, \qquad S_0^\pi = \varnothing, \tag{1}$$

define a complete accuracy path rather than a single subset. For $S_k = S_{k-1} \cup \{m_{\pi_k}\}$, decompose finite-sample risk into Bayes and excess components:

$$R_n(S) = R^\star(S) + E_n(S), \quad g_k := R^\star(S_{k-1}) - R^\star(S_k) \geq 0, \quad c_k := E_n(S_k) - E_n(S_{k-1}). \tag{2}$$

The one-step identity $R_n(S_k) - R_n(S_{k-1}) = c_k - g_k$ shows that an added view helps exactly when its fresh task signal exceeds its cost. Under log loss,

$$g_k = I\left(F_{m_{\pi_k}}; Y \mid \bar{F}_{S_{k-1}}\right), \tag{3}$$

so redundancy reduces later gains and makes the peak order-dependent.

Theorem 1 (Order-specific gain–cost crossing). *If g_k is non-increasing, c_k is non-decreasing, and $g_k - c_k$ changes from positive to negative with non-zero margin, then risk decreases up to the unique crossing and increases thereafter. Because $\{g_k, c_k\}$ depends on π, both peak location and height are order-specific.*

Proposition 1 (Peak lift by front-loading). *Writing* $a_k^\pi = g_k^\pi - c_k^\pi$, $\min_k R_n(S_k^\pi) = R_n(S_0^\pi) - \max_k \sum_{i=1}^k a_i^\pi$*; hence front-loading positive marginal utility raises the best prefix.*

Full proofs and failure-mode constructions are in Appendix A.3.

2.2 Diagnostic evidence

We test a fixed, favourable six-encoder trajectory (CLIP, DINOv2, ConvNeXt, SigLIP, MAE, ResNet-50) with gated fusion (Shi et al., 2025; Tong et al., 2024). The five datasets map one-to-one to generic, fine-grained, texture, geospatial, and structured-symbol regimes. Table 1 summarises both the 10-shot decline and how the peak moves with sample size; full prefix curves and configurations are in Appendix A.3.

At 10-shot, the full six-view prefix loses 3.92 points on Country211 and 8.57 on GTSRB relative to the best prefix; the optimum is therefore a property of the ordered utility sequence, not model count. As n grows, the peak moves right on CIFAR-100, Pets, and GTSRB, stays at one on single-encoder-saturated Country211, and is non-monotonic on DTD. This motivates a selector that front-loads task-aligned views.

2.3 Why additional views hurt

The one-step identity separates two mechanisms. **Type I (task-conditional redundancy)** occurs when $I(F_{m_k}; Y \mid \bar{F}_{S_{k-1}}) = 0$, so $g_k = 0$ under log loss and any positive excess-risk increment c_k worsens performance. **Type II (task-model misalignment)** permits a candidate to be representationally diverse yet have $g_k \le \epsilon$; whenever $c_k > \epsilon$, the step increases risk by at least $c_k - \epsilon$. Thus task-agnostic diversity can prefer an encoder that is novel but not useful. Appendix A.3.6 provides the formal definitions and proofs.

3 KAGES: Greedy View-Set Composition by Kernel Alignment

Section 2 showed that large-pool fusion is an ordering problem: a selector's quality is reflected in the entire prefix-accuracy curve, not in a single best prefix. KAGES constructs such an ordering by greedily maximising a set-level kernel-alignment utility. It is training-free but not label-free: it uses labels only through a label kernel and never trains a downstream classifier for a candidate subset, which makes it deployable at pool sizes where validation-based subset search is infeasible.

3.1 Set-level objective and greedy selection

Let $\mathcal{M} = \{m_1, \dots, m_M\}$ be the candidate pool, $F_s \in \mathbb{R}^{n \times d_s}$ the feature matrix of encoder s on n labelled samples, and $Y \in \{0,1\}^{n \times C_{\text{cls}}}$ the one-hot label matrix. For a subset $S \subseteq \mathcal{M}$, KAGES assigns the set-level score

$$J(S) \;=\; U(\bar{F}_S, Y) \;-\; \tau \sum_{s \in S} \frac{d_s}{n}, \tag{4}$$

where $U(\bar{F}_S, Y) \in \mathbb{R}$ measures the task alignment of the joint selected representation $\bar{F}_S$ and the optional second term is a per-encoder finite-sample complexity penalty ($\tau \ge 0$; the default $\tau = 0$ recovers the pure utility and is swept in Section 4). We set $U(\emptyset, Y) := J(\emptyset) := 0$ by convention. Given a current set S, KAGES scores a candidate $m \notin S$ by the marginal gain $\Delta J(m \mid S) := J(S \cup \{m\}) - J(S)$, adds the candidate with the largest marginal gain, and repeats until all candidates are ordered. The output is a *complete* encoder ordering; downstream users may select any prefix according to their compute budget. When a single score-based deployment size is desired, an optional rule is $\hat{k}_J := \min\big(\arg\max_{1 \le k \le M} J(S_k)\big)$, which chooses the smallest prefix attaining the largest score along the completed path.

Algorithm 1 produces a complete ordering of the pool and, optionally, a score-selected prefix. The algorithm itself does not require monotonicity or submodularity; Theorem 2 applies only to the unpenalised alignment utility and only under its stated conditions.

Table 1: **Order-specific peaks across regimes.** $k^\star$ is the best prefix at 10-shot, 100-shot, and full data. "All" is 10-shot accuracy after all six encoders; Gap is Peak−All.

Dataset	$k^\star_{10}$	$k^\star_{100}$	$k^\star_{\text{full}}$	Peak_{10}	All_{10}	Gap
CIFAR-100	3	3	5	80.37	79.92	+0.45
Pets	3	4	5	93.75	93.49	+0.26
DTD	6	3	5	74.18	74.18	0.00
Country211	1	1	1	13.50	9.58	+3.92
GTSRB	1	1	5	61.15	52.58	+8.57

Table 2: **10-shot accuracy (%) along the six-encoder diagnostic path.** Bold marks the best prefix; $\star$ denotes tasks where the first encoder remains optimal.

Dataset	1	2	3	4	5	6	Peak
CIFAR-100	60.29	80.11	**80.37**	80.18	80.20	79.92	3
Pets	85.04	92.86	**93.75**	93.46	93.65	93.49	3
DTD	64.41	72.67	73.52	74.05	74.17	**74.18**	6
Country211$\star$	**13.50**	9.85	9.91	9.82	9.71	9.58	1
GTSRB$\star$	**61.15**	51.54	52.73	52.34	51.97	52.58	1

3.2 Centred kernel-target alignment as utility

For non-empty S, we instantiate U with the centred set-level kernel-target alignment,

$$U(S) = \text{CKA}(K^c_S, K^c_Y), \quad K^c_S = \sum_{s\in S} K^c_s, \quad K^c_s = HF_sF_s^\top H, \quad K^c_Y = HYY^\top H, \tag{5}$$

where $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix and K^c_s is the centred linear kernel of encoder s. By the convention above, $U(\emptyset, Y) = 0$; we assume $\|K^c_Y\|_F > 0$ and discard any encoder with $\|K^c_s\|_F = 0$. For non-zero kernels, centred kernel-target alignment (Cristianini et al., 2002; Kornblith et al., 2019) is bounded in $[0, 1]$, scale-invariant in each kernel individually, and reduces to canonical kernel-target alignment for $|S| = 1$.

Linear-kernel additivity means K^c_S is exactly the centred kernel of concatenated features (derivation in Appendix A.3). KAGES therefore updates one $n \times n$ running kernel and evaluates each candidate with Frobenius inner products in $\mathcal{O}(n^2)$ time, independent of encoder dimension. A full traversal of 10 encoders at $n = 500$ takes under one second on one GPU. We use raw additive kernels; the per-encoder normalised ablation is reported in Appendix A.3.10.

3.3 Algebraic structure and greedy guarantee

For the pure utility used in the main results ($\tau = 0$), additivity gives an exact relevance–redundancy decomposition.

Proposition 2 (Algebraic expansion of $\text{CKA}(K^c_S, K^c_Y)$)**.** *For any non-empty S with non-zero set and label kernels,*

$$\text{CKA}\big(K^c_S,\, K^c_Y\big) \;=\; \frac{\displaystyle\sum_{s\in S} \langle K^c_s, K^c_Y\rangle_F}{\displaystyle\sqrt{\sum_{s,s'\in S} \langle K^c_s, K^c_{s'}\rangle_F}\;\cdot\; \|K^c_Y\|_F}. \tag{6}$$

The numerator accumulates target relevance; the denominator also contains pairwise between-encoder similarity.

Exact marginal-gain criterion. Write $A_S = \langle K^c_S, K^c_Y\rangle_F$, $B_S = \|K^c_S\|^2_F$, $r_m = \langle K^c_m, K^c_Y\rangle_F$, $h_m = \langle K^c_m, K^c_S\rangle_F$, and $q_m = \|K^c_m\|^2_F$.

Algorithm 1 KAGES: training-free greedy encoder ordering by set-level marginal utility.

Require: Pool $\mathcal{M} = \{m_1, \ldots, m_M\}$; labelled task (X, Y) with n samples; cost coefficient $\tau \geq 0$.
1: Compute centred encoder kernels K_s^c for every $s \in \mathcal{M}$ on the n samples.
2: Initialize $S_0 \leftarrow \emptyset$, $\hat{\pi} \leftarrow (\,)$.
3: **for** $t = 1, 2, \ldots, M$ **do**
4: $\quad m_t \leftarrow \arg\max_{m \notin S_{t-1}} \Delta J(m \mid S_{t-1})$ *# marginal-gain greedy step*
5: $\quad S_t \leftarrow S_{t-1} \cup \{m_t\}$; append m_t to $\hat{\pi}$.
6: **end for**
7: Optional: $\hat{k}_J \leftarrow \min\big(\arg\max_{1 \leq k \leq M} J(S_k)\big)$.
8: **return** encoder ordering $\hat{\pi} = (m_1, \ldots, m_M)$ and optional score-selected prefix $S_{\hat{k}_J}$.

Proposition 3 (Exact CKA marginal-gain criterion)**.** *For non-empty S with $B_S > 0$, adding $m \notin S$ improves alignment iff*

$$r_m \;>\; A_S \cdot \left(\sqrt{1 + \frac{2h_m + q_m}{B_S}} \;-\; 1 \right). \tag{7}$$

Thus target relevance r_m must exceed an adaptive penalty determined by overlap h_m, candidate mass q_m, and the current set. When the candidate contributions are small relative to the selected set, first-order expansion gives

$$\Delta U(m \mid S) \approx \frac{1}{\sqrt{B_S C}} \left[r_m - \frac{A_S}{2B_S}(2h_m + q_m) \right], \tag{8}$$

where $C = \|K_Y^c\|_F^2$. KAGES therefore implements a parameter-free kernel mRMR criterion whose redundancy weight $A_S/(2B_S)$ adapts to the current set rather than being tuned globally (Peng et al., 2005; Brown et al., 2012). Appendix A.3.4 gives the full derivation.

Conditional greedy guarantee. Normalised CKA need not be monotone, so the guarantee is explicitly conditional rather than universal.

Theorem 2 (Conditional prefix-wise guarantee)**.** *Let $U(\emptyset) = 0$ and suppose U is monotone. If, up to budget K, its submodularity ratio is at least $\gamma \in (0, 1]$ (Das & Kempe, 2011), i.e., for every admissible L, T,*

$$\sum_{m \in T} \Delta U(m \mid L) \geq \gamma \, [U(L \cup T) - U(L)],$$

then every greedy prefix satisfies

$$U(S_k^{\pi_g}) \;\geq\; \big(1 - e^{-\gamma}\big) \max_{|S| \leq k} U(S).$$

The proof contracts the optimality gap by at least a γ/k fraction at each step; Appendix A.1 gives the complete argument and a mutual-information specialisation. This is a fixed-budget utility guarantee, not a claim that empirical CKA is always monotone or that the score-selected stopping point is accuracy-optimal.

Empirical ordering consistency. KAGES uses empirical rather than population marginal gains. Under the uniform subgaussian concentration, bias, and positive path-margin assumptions in Appendix A.2, if $n \geq (4B/G)^2$ then its ordering matches the population greedy path with probability at least $1 - [M(M+1) - 2]\exp[-nG^2/(32\sigma^2)]$. The statement is assumption-conditional and is not an unconditional concentration claim for empirical CKA.

4 Experiments

4.1 Experimental setup

Datasets and encoder pool. The five datasets cover generic objects (CIFAR-100), fine-grained recognition (Oxford Pets), texture/material (DTD), geospatial scenes (Country211), and struc-

tured symbols (GTSRB). The main pool contains 10 frozen backbones: CLIP and SigLIP (language supervision) (Radford et al., 2021; Zhai et al., 2023); DINOv2, MAE, and Data2Vec (self-supervision) (Oquab et al., 2024; He et al., 2022; Baevski et al., 2022); Stable-Diffusion VAE (generative) (Rombach et al., 2022); ConvNeXt, ResNet-50, and ViT-B/16 (class supervision) (Liu et al., 2022; He et al., 2016; Dosovitskiy et al., 2021); and SAM-ViT-B (task expert) (Kirillov et al., 2023). The latter two paradigms serve as deliberately difficult candidates that a task-aware selector should defer.

Protocols and metrics. Protocol **B** uses 16-shot training with a Concat+MLP fusion head; **H** uses 8 shots and a larger $M = 15$ candidate pool; and **F** uses the full training set with the same head family. Together they test low-data, larger-pool, and full-data behaviour. Because view-set composition produces an ordering, the primary metric is **AULC**, the mean accuracy over all prefixes. We also report **Acc@**$k^\star$, the best downstream accuracy reached along the path. This is a hindsight diagnostic and is not presented as an evaluation of the optional score-selected stopping rule.

Baselines and reporting. We compare the standard full-fusion order (Fuse-All), random ordering, CKA diversity, DPP MAP inference, and facility-location selection. Results average seeds $\{17, 23, 42, 71, 137\}$. Appendix A.3.1 gives training details and additional results.

4.2 Main results

KAGES improves the entire prefix path. It is best against Fuse-All, random, and diversity in every cell of Table 3, leading Fuse-All by 3.9, 5.8, and 3.3 average AULC points on B, H, and F. The largest gains occur on GTSRB ($+8.9$ to $+15.5$), where task alignment matters most. KAGES also exceeds DPP and facility-location selection on average in every protocol (B: 66.93 vs $65.21/66.13$; H: 64.77 vs $62.13/63.49$; F: 74.78 vs $73.51/73.98$). Figure 2 shows that these gains arise because useful encoders are selected early.

Selection is interpretable and robust. DINOv2 is first on all seeds for the four natural-image regimes, while GTSRB selects ResNet-50 or CLIP; the generative and task-expert encoders are never first. A cost sweep $\tau \in \{0, .01, .05, .1, .5, 1\}$ gives mean AULC $\{66.9, 66.7, 65.5, 64.6, 62.9, 62.5\}$, so the pure alignment objective ($\tau = 0$) is retained. Additional experimental details are in Appendix A.3.

4.3 Stronger and single-model baselines

KAGES remains best when the stronger DPP and facility-location baselines are shown per dataset, not only in aggregate. The published 16-shot adapter results provide context rather than a controlled comparison: on DTD and Flowers the fused representation exceeds recent CLIP adapters, while GTSRB shows the largest benefit because its structured-symbol regime requires a task-aligned model subset. [†]The average uses the three datasets available for the multi-model rows; missing single-model results are not imputed.

4.4 Sensitivity and stability

The penalty increasingly tracks encoder dimension rather than task utility, explaining the monotonic degradation in Table 5. Per-encoder Frobenius normalisation also lowers AULC by approximately 0.4 points, so the raw additive kernel is retained. Across the five seeds, KAGES's per-dataset AULC standard deviation is at most 0.3 points in the reported aggregate summary; variability and normalisation summaries are in Appendices A.3.10 and A.3.11.

4.5 Beyond classification

The effect extends beyond classification. In image retrieval, Recall@1 peaks later ($k = 8$ on Pets and $k = 10$ on DTD/GTSRB), and KAGES reaches each peak. Treating six frozen LLMs as views, KAGES reaches 75.0% on MMLU at $k = 2$ and 91.0% on ARC-Challenge at $k = 4$, attaining the peak with fewer models than fixed or diversity-greedy orders (Figure 3). Full curves, CKA matrices, and protocol details are in Appendices A.3.7 and A.4.

Table 3: **AULC (%) across three protocols.** KAGES is best among the multi-view defaults in every cell. Means over five seeds; bold marks the best value per protocol and dataset.

Method	C100	Pets	DTD	C211	GTSRB	Avg
Protocol B (16-shot, Concat+MLP)						
Fuse-All	78.54	94.06	75.35	10.16	57.18	63.06
Random	77.06	92.97	74.87	11.10	59.16	63.03
Diversity	80.61	90.91	74.38	11.27	59.12	63.26
KAGES (ours)	**83.84**	**94.70**	**77.82**	**12.16**	**66.12**	**66.93**
Δ *vs Fuse-All*	*+5.3*	*+0.6*	*+2.5*	*+2.0*	*+8.9*	*+3.9*
Protocol H (8-shot, M=15)						
Fuse-All	73.89	89.01	70.10	9.65	52.42	59.02
Random	75.11	91.78	69.84	9.77	54.88	60.28
Diversity	78.18	86.91	68.87	9.67	56.67	60.06
KAGES (ours)	**80.20**	**92.82**	**72.36**	**10.57**	**67.92**	**64.77**
Δ *vs Fuse-All*	*+6.3*	*+3.8*	*+2.3*	*+0.9*	*+15.5*	*+5.8*
Protocol F (full data)						
Fuse-All	85.07	95.90	80.05	16.51	79.71	71.45
Random	84.66	95.26	80.07	19.43	81.87	72.26
Diversity	86.21	94.35	79.85	18.51	81.31	72.05
KAGES (ours)	**88.19**	**96.25**	**82.96**	**20.07**	**86.41**	**74.78**
Δ *vs Fuse-All*	*+3.1*	*+0.4*	*+2.9*	*+3.6*	*+6.7*	*+3.3*

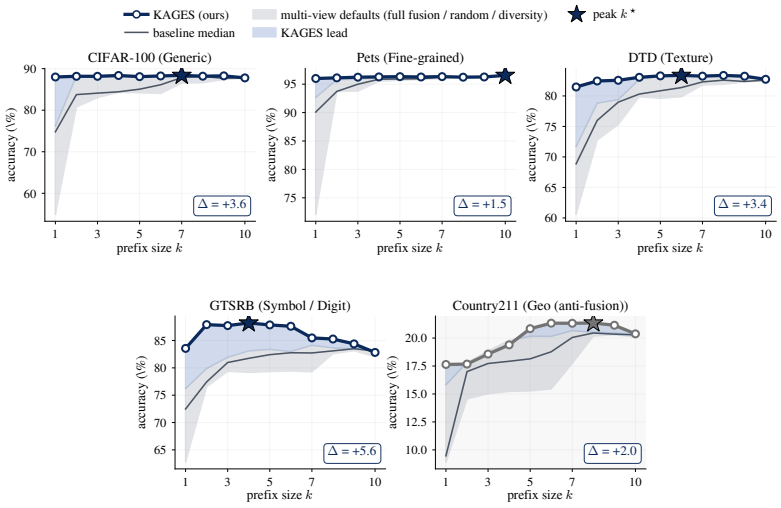

Figure 2: **Prefix accuracy under Protocol F.** KAGES (navy; star at peak) front-loads useful encoders and stays near the top of the gray baseline envelope.

5 Related Work

Fusion and model selection. Multi-view learning studies small fixed view sets (Xu et al., 2013; Yan et al., 2021; Trosten et al., 2021); multi-encoder systems such as Eagle, Cambrian-1, BRAVE, and MoVA extend fusion to pretrained backbones but retain relatively small prescribed pools (Shi et al., 2025; Tong et al., 2024; Kar et al., 2024; Zong et al., 2024). Transferability scores including LogME, H-Score, GBC, LEEP, and PACTran rank encoders independently (You et al., 2021; Bao et al., 2019; Pandy et al., 2022; Nguyen et al., 2020; Ding et al., 2022), while DPP and facility-location methods reward diversity or coverage. KAGES instead scores each candidate conditional on the already-selected set and constructs a complete ordering.

Kernel selection. mRMR/JMI and kernel feature selectors combine relevance and redundancy (Peng et al., 2005; Yang & Moody, 1999; Brown et al., 2012; Song et al., 2012; Yamada et al., 2014),

Table 4: **Additional comparisons.** Top: complete Protocol-B AULC. Bottom: comparison with published 16-shot single-model CLIP-adaptation results. Bold/underline mark best/second-best.

Method	C100	Pets	DTD	C211	GTSRB	Avg
Fuse-All	78.54	<u>94.06</u>	75.35	10.16	57.18	63.06
Diversity	80.61	90.91	74.38	11.27	59.12	63.26
DPP	<u>83.62</u>	93.21	72.48	<u>11.87</u>	64.88	65.21
Facility location	82.82	93.31	<u>77.64</u>	11.34	<u>65.53</u>	<u>66.13</u>
KAGES	**83.84**	**94.70**	**77.82**	**12.16**	**66.12**	**66.93**

Method (16-shot)	DTD	Flowers	GTSRB	Avg†
CoOp (Zhou et al., 2022)	68.7	95.7	–	–
Tip-Adapter-F (Zhang et al., 2022)	73.7	97.2	–	–
TaskRes (Yu et al., 2023)	73.4	97.0	–	–
CaFo (Zhang et al., 2023)	74.6	97.5	–	–
All Models	77.9	99.6	61.6	79.7
KAGES	**78.7**	**99.7**	**72.1**	**83.5**

Table 5: **Cost-penalty sensitivity on Protocol B.** Mean AULC is highest for the pure alignment objective.

τ	0	.01	.05	.1	.5	1.0
Mean AULC (%)	**66.9**	66.7	65.5	64.6	62.9	62.5

while CKA/HSIC measure representation similarity (Gretton et al., 2005; Kornblith et al., 2019). KAGES treats an entire encoder representation as one atomic view: its exact positive-gain criterion uses an adaptive redundancy weight determined by the current set, costs $\mathcal{O}(n^2)$ independent of feature dimension, and provides a conditional guarantee for every prefix rather than one fixed budget. Appendix A.5 gives the extended comparison.

6 Discussion, Limitations, and Future Work

Limitations. The normalised CKA objective is neither generally monotone nor generally submodular because the denominator in Eq. 6 can outweigh a non-negative numerator increment and introduces cross-overlap terms. Our $(1 - e^{-\gamma})$ guarantee is therefore conditional on monotonicity and a positive submodularity ratio; establishing when empirical frozen-encoder pools satisfy these conditions is future work. KAGES never first-picks the generative and task-expert candidates; these selection results alone do not establish their downstream fusion benefit.

Scope and computation. KAGES assumes a fixed pool of frozen encoders and a labelled sample from the target task; it is training-free during selection but not label-free. The $n \times n$ kernels give encoder-dimension-independent scoring but impose quadratic time and memory in the number of selection samples. Our largest pool has $M = 15$, and the non-vision evidence uses linear probing rather than end-to-end multi-LLM fusion. These boundaries prevent us from claiming that the same ordering remains optimal after encoder fine-tuning, for dynamically growing pools, or for generative decoding objectives.

Stopping and future work. The experiments evaluate complete prefix paths through AULC and hindsight peak accuracy. Although $J(S_k)$ supplies an optional stopping rule, we do not claim that its maximiser is downstream-accuracy optimal. Calibrating that rule, approximating kernels for much larger n, and updating the ordering under streaming labels or newly released encoders are natural next steps; Appendix A.6 sketches the corresponding sequential formulation.

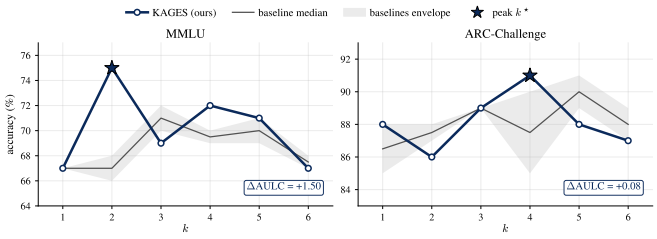

Figure 3: **Frozen-LLM hidden-state linear probing.** KAGES reaches the MMLU and ARC-Challenge peaks with fewer models than fixed-order and diversity-greedy baselines.

Table 6: **Cross-modal peak summary.** Optimal prefix size is task-dependent: retrieval benefits from larger sets, while LLM fusion peaks early.

Setting	Metric	Peak k	Best (%)
DTD retrieval	Recall@1	10	74.6
GTSRB retrieval	Recall@1	10	60.2
Pets retrieval	Recall@1	8	94.3
MMLU-500	Accuracy	2	75.0
ARC-Ch-500	Accuracy	4	91.0

7 Conclusion

Large foundation-model pools make view selection as important as fusion. KAGES orders frozen encoders by set-level kernel alignment, front-loads task-relevant views, and improves prefix-path quality across regimes and data scales without downstream retraining during selection. Its guarantees remain conditional on explicit monotonicity and submodularity-ratio assumptions; the experiments establish the practical benefit. The results support a simple conclusion: effective large-pool fusion depends on choosing a compact, complementary set rather than using every available encoder.

AI use statement

Generative AI tools were used to improve manuscript readability, suggest a more concise paper structure, and assist with LaTeX formatting. All AI-assisted edits were reviewed against the source manuscript and reported results. The tools were not used in this format-conversion process to generate datasets, run experiments, or alter numerical results. The authors take responsibility for the final content, including AI-assisted text and formatting.

Reproducibility statement

The appendix provides experimental details, aggregate results, ablation summaries, cross-modal evaluations, and proofs. The accompanying source package includes all tables, figures, bibliography entries, and the official ICLR 2027 style files needed to rebuild this document.

REFERENCES

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, pp. 1298–1312, 2022.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *ICIP*, pp. 2309–2313, 2019.

Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *JMLR*, 13:27–66, 2012.

Yimeng Chen, Tianyang Hu, Fengwei Zhou, Zhenguo Li, and Zhiming Ma. Explore and exploit the diverse knowledge in model zoo for domain generalization. In *ICML*, pp. 4623–4640, 2023.

N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. *Advances in Neural Information Processing Systems (NeurIPS)*, 2002.

A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *International Conference on Machine Learning (ICML)*, 2011.

Nan Ding, Xi Chen, Tomer Levinboim, Beer Changpinyo, and Radu Soricut. PACTran: PAC-Bayesian metrics for estimating the transferability of pretrained models to classification tasks. In *ECCV*, 2022. arXiv:2203.05126.

Qishi Dong, Awais Muhammad, Fengwei Zhou, Chuanlong Xie, Tianyang Hu, Yongxin Yang, Sung-Ho Bae, and Zhenguo Li. ZooD: Exploiting model zoo for out-of-distribution generalization. In *NeurIPS*, 2022.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Abhimanyu Dubey et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *COLT*, 2002.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *JMLR*, 7:1079–1105, 2006.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, pp. 63–77, 2005.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

Long-Kai Huang, Ying Wei, Yu Rong, Qiang Yang, and Junzhou Huang. Frustratingly easy transferability estimation. In *ICML*, 2022. arXiv:2106.09362.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024.

Oğuzhan Fatih Kar, Alessio Tonioni, et al. BRAVE: Broadening the visual encoding of vision-language models. In *ECCV*, 2024.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *JMLR*, 17(1), 2016.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019.

Yuchen Liu, Yaoming Zhu, Jiangbo Wang, Enwei Yu, Junxiang Liu, and Junfeng Yan. From CLIP to DINO: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pp. 11976–11986, 2022.

Fanqing Meng, Wenqi Shao, Zhanglin Peng, Chonghe Jiang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Foundation model is efficient multimodal multitask model selector. In *NeurIPS*, 2023. arXiv:2308.06262.

Cuong V Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *ICML*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

Michal Pandy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using bhattacharyya class separability. In *CVPR*, pp. 9162–9172, 2022.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE TPAMI*, 27(8):1226–1238, 2005.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *ICLR*, 2025.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012.

Yang Tan, Yang Li, and Shao-Lun Huang. OTCE: A transferability metric for cross-domain cross-task representations. In *CVPR*, 2021. arXiv:2103.13843.

Yang Tan, Enming Zhang, Yang Li, Shao-Lun Huang, and Xiao-Ping Zhang. Transferability-guided cross-domain cross-task transfer learning. *IEEE TNNLS*, 2024. arXiv:2207.05510.

Alexander Tartakovsky, Igor Nikiforov, and Michèle Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, 2014.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *NeurIPS*, 2024.

Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *ICCV*, pp. 1395–1405, 2019.

Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *CVPR*, 2021.

KB Vimal, Saketh Bachu, Tanmay Garg, Niveditha Lakshmi Narasimhan, Raghavan Konuru, and Vineeth N Balasubramanian. Building a winning team: Selecting source model ensembles using a submodular transferability estimation approach. In *ICCV*, pp. 11575–11586, 2023.

Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

Makoto Yamada, Wittawat Jitkrittum, Leonid Sigal, Eric P. Xing, and Masashi Sugiyama. High-dimensional feature selection by feature-wise kernelized Lasso. *Neural Computation*, 26(1): 185–207, 2014.

Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Howard Hua Yang and John Moody. Data visualization and feature selection: New algorithms for nongaussian data. In *NeurIPS*, pp. 687–693, 1999.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *ICML*, 2021.

Kaichao You, Yong Liu, Ziyang Zhang, Jianmin Wang, Michael I. Jordan, and Mingsheng Long. Ranking and tuning pre-trained models: A new paradigm for exploiting model hubs. *JMLR*, 23 (209), 2022. arXiv:2110.10545.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *CVPR*, pp. 10899–10909, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11975–11986, 2023.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 2022.

Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *CVPR*, pp. 15211–15222, 2023.

Tengxue Zhang, Yang Shu, Xinyang Chen, Yifei Long, Chenjuan Guo, and Bin Yang. Assessing pre-trained models for transfer learning through distribution of spectral components. In *AAAI*, 2025. arXiv:2412.19085.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022.

Zhuofan Zong et al. MoVA: Adapting mixture of vision experts to multimodal context. In *NeurIPS*, 2024.

APPENDIX OVERVIEW

The appendix supports specific portions of the main paper:

- **App. A.1:** conditional $(1 - e^{-\gamma})$ greedy guarantee under a standard submodularity-ratio condition. A mutual-information specialization recovers the $(1 - 1/e)$ bound.
- **App. A.2:** ordering-consistency proof for ΔJ (Theorem A.2), supporting the empirical-ordering claim at the end of Section 3.3 of the main paper.
- **App. A.3:** experimental configuration, additional results tables, risk decomposition, and proof of the conditional marginal-utility crossing (Theorem 1). Supports Sections 2 and 4 of the main paper.
- **App. A.4:** cross-modal validation on frozen LLMs (supports Section 4.5).
- **App. A.5:** an extended comparison to foundation-model fusion, transferability scoring, and kernel selection.
- **App. A.6:** future work, online and sequential extensions.

The notation of the main paper carries over: $\mathcal{M}$ is the encoder pool of size M, and $\hat{\pi}$ is the KAGES greedy ordering on the labelled task, as defined in Section 3 of the main paper.

A.1 GREEDY APPROXIMATION GUARANTEE UNDER A SUBMODULARITY RATIO

We give the conditional fixed-budget greedy guarantee used in the main paper. The result is stated for a generic non-negative set utility U and applies only when the stated monotonicity and submodularity-ratio conditions hold. We then give a population mutual-information specialization for which these conditions follow from conditional independence.

For $m \notin S$, write $\Delta U(m \mid S) := U(S \cup \{m\}) - U(S)$.

Theorem A.1 (Conditional greedy guarantee under a submodularity ratio)**.** *Let $U : 2^{\mathcal{M}} \to \mathbb{R}_{\geq 0}$ satisfy $U(\emptyset) = 0$ and be monotone non-decreasing. Fix a cardinality budget K. Suppose there is a constant $\gamma \in (0, 1]$ such that, for every $L \subseteq \mathcal{M}$ and every $T \subseteq \mathcal{M} \setminus L$ with $|T| \leq K$,*

$$\sum_{m \in T} \Delta U(m \mid L) \;\geq\; \gamma\big[U(L \cup T) - U(L)\big]. \tag{A.1}$$

Let $S_0 = \emptyset$ and let greedy choose

$$m_k \in \arg\max_{m \notin S_{k-1}} \Delta U(m \mid S_{k-1}), \qquad S_k = S_{k-1} \cup \{m_k\}.$$

Then

$$U(S_K) \;\geq\; \big(1 - e^{-\gamma}\big) \max_{|S| \leq K} U(S).$$

Proof. Let $O \in \arg\max_{|S| \leq K} U(S)$ and write $G_k := U(O) - U(S_k)$. By monotonicity and (A.1),

$$\sum_{m \in O \setminus S_k} \Delta U(m \mid S_k) \;\geq\; \gamma\big[U(S_k \cup O) - U(S_k)\big] \;\geq\; \gamma G_k.$$

Since $|O \setminus S_k| \leq K$, the largest available marginal gain is at least $\gamma G_k / K$. Greedy therefore satisfies

$$U(S_{k+1}) - U(S_k) \geq \frac{\gamma}{K} G_k, \qquad G_{k+1} \leq \left(1 - \frac{\gamma}{K}\right) G_k.$$

Iterating from $S_0 = \emptyset$ gives

$$G_K \leq \left(1 - \frac{\gamma}{K}\right)^K U(O) \leq e^{-\gamma} U(O),$$

which proves the claim. □

Corollary A.1 (Prefix-wise fixed-budget guarantee)**.** *Under the hypotheses of Theorem A.1, for every $k = 1, \ldots, K$, the same greedy ordering satisfies*

$$U(S_k) \geq (1 - e^{-\gamma}) \max_{|S| \leq k} U(S).$$

Proof. Apply Theorem A.1 with the budget set to k. The first k greedy choices are unchanged when the traversal is subsequently continued to budget K. □

For the information-theoretic specialization, let

$$\bar{F}_S := (F_m)_{m \in S}$$

denote the joint collection of the selected encoder features.

Lemma A.1 (Conditional independence implies information submodularity)**.** *The set function $f(S) := I(\bar{F}_S; Y)$ is normalized and monotone non-decreasing. If the encoder features are mutually independent conditional on Y, then f is submodular and hence satisfies* (A.1) *with $U = f$ and $\gamma = 1$.*

Proof. Normalization is immediate. Monotonicity follows from the chain rule,

$$f(A \cup \{m\}) - f(A) = I(F_m; Y \mid \bar{F}_A) \geq 0.$$

For $A \subseteq B$ and $m \notin B$,

$$I(F_m; Y \mid \bar{F}_A) = H(F_m \mid \bar{F}_A) - H(F_m \mid Y, \bar{F}_A).$$

Conditional independence gives $H(F_m \mid Y, \bar{F}_A) = H(F_m \mid Y)$, and similarly for B. Since conditioning cannot increase entropy, $H(F_m \mid \bar{F}_A) \geq H(F_m \mid \bar{F}_B)$, so

$$I(F_m; Y \mid \bar{F}_A) \geq I(F_m; Y \mid \bar{F}_B).$$

This is the diminishing-returns condition for submodularity. A normalized monotone submodular function satisfies (A.1) with $\gamma = 1$. □

Corollary A.2 ($(1-1/e)$ information-gain guarantee)**.** *Under the conditional-independence condition of Lemma A.1, greedy maximization of $f(S) = I(\bar{F}_S; Y)$ satisfies*

$$f(S_K^{\text{greedy}}) \geq (1 - 1/e) \max_{|S| \leq K} f(S).$$

Scope of the guarantee. Theorem A.1 is conditional: it does not assert that every frozen-encoder pool, or the empirical CKA objective used by KAGES, is automatically monotone or has a positive submodularity ratio. Because CKA contains a normalization denominator, non-negative target alignment of a candidate kernel alone does not guarantee a non-negative marginal CKA gain; the exact one-step condition is derived in App. A.3.4. The theorem and Corollary A.1 provide fixed-budget prefix guarantees only when the stated conditions hold.

A.2 Ordering Consistency for ΔJ

KAGES (Algorithm 1 in the main paper) acts on the empirical utility $\widehat{J}_n$ computed from n labelled samples. This appendix shows that, under uniform subgaussian concentration of the marginal gains and a non-vanishing per-step population margin, empirical KAGES recovers the population greedy ordering with exponentially high probability.

A.2.1 Setup and notation

Let $\mathcal{M} = \{1, \ldots, M\}$ be a fixed encoder pool with $M \geq 2$. Write $J(S)$ for a deterministic population target utility (for example, a population limit of the alignment score when such a limit exists) and $\widehat{J}_n(S)$ for its n-sample empirical estimator. The theorem below is generic and relies only on the stated concentration and bias assumptions; it does not derive those assumptions for empirical CKA. Marginal gains are

$$\Delta J(m \mid S) := J(S \cup \{m\}) - J(S), \qquad \Delta \widehat{J}_n(m \mid S) := \widehat{J}_n(S \cup \{m\}) - \widehat{J}_n(S).$$

The population greedy trajectory is $(m_t^\star, S_t^\star)_{t=1}^M$ with $S_0^\star = \emptyset$ and

$$m_t^\star \in \arg\max_{m \notin S_{t-1}^\star} \Delta J(m \mid S_{t-1}^\star), \qquad S_t^\star = S_{t-1}^\star \cup \{m_t^\star\}.$$

The empirical trajectory $(\hat{m}_t, \hat{S}_t)$ is defined analogously using $\Delta \widehat{J}_n$. Ties are resolved by the same fixed deterministic rule in both trajectories.

A.2.2 ASSUMPTIONS

Assumption A.1 (Sample independence and pool finiteness)**.** *The n labelled observations are i.i.d. from a fixed joint distribution, and the pool cardinality M is fixed independently of n.*

Assumption A.2 (Uniform subgaussian concentration with bias)**.** *For every $m \in \mathcal{M}$ and every $S \subseteq \mathcal{M} \setminus \{m\}$,*

$$\Pr\Big[\Big|\Delta\widehat{J}_n(m \mid S) - \mathbb{E}\,\Delta\widehat{J}_n(m \mid S)\Big| > t\Big] \le 2\exp\left(-\frac{nt^2}{2\sigma^2}\right) \quad \textit{for all } t > 0,$$

with $\sigma > 0$. Moreover, the bias

$$b_n(m, S) := \mathbb{E}\,\Delta\widehat{J}_n(m \mid S) - \Delta J(m \mid S)$$

satisfies $|b_n(m, S)| \le B/\sqrt{n}$ uniformly over (m, S).

Assumption A.3 (Greedy ordering margin)**.** *For every step $t \in \{1, \dots, M-1\}$,*

$$\Delta J(m_t^\star \mid S_{t-1}^\star) - \max_{m \notin S_{t-1}^\star \cup \{m_t^\star\}} \Delta J(m \mid S_{t-1}^\star) \ge G$$

for some $G > 0$.

A.2.3 THEOREM STATEMENT

Theorem A.2 (Ordering consistency for ΔJ)**.** *Under Assumptions A.1–A.3, if $n \ge (4B/G)^2$, then*

$$\Pr\big[\hat{\pi} \ne (m_1^\star, \dots, m_M^\star)\big] \le \big[M(M+1) - 2\big]\exp\left(-\frac{nG^2}{32\sigma^2}\right).$$

Consequently, a target failure probability $\epsilon \in (0, 1)$ is attained whenever

$$n \ge \max\left\{\left(\frac{4B}{G}\right)^2, \frac{32\sigma^2}{G^2}\log\frac{M(M+1) - 2}{\epsilon}\right\}.$$

In particular, the sufficient sample size is $O\big((B^2 + \sigma^2[\log M + \log(1/\epsilon)])/G^2\big)$.

A.2.4 PROOF

Step 1: deviation from the population marginal gain. If $n \ge (4B/G)^2$, then $|b_n(m, S)| \le G/4$. Hence, for any fixed (m, S),

$$\begin{aligned}&\Pr\left[\Big|\Delta\widehat{J}_n(m \mid S) - \Delta J(m \mid S)\Big| \ge \frac{G}{2}\right] \\ &\quad\le \Pr\left[\Big|\Delta\widehat{J}_n(m \mid S) - \mathbb{E}\,\Delta\widehat{J}_n(m \mid S)\Big| \ge \frac{G}{4}\right] \le 2\exp\left(-\frac{nG^2}{32\sigma^2}\right).\end{aligned}$$

Step 2: deterministic good events along the population path. For every population-path step $t \in \{1, \dots, M-1\}$ and every candidate $m \notin S_{t-1}^\star$, define

$$\mathcal{E}_{t,m} := \left\{\Big|\Delta\widehat{J}_n(m \mid S_{t-1}^\star) - \Delta J(m \mid S_{t-1}^\star)\Big| < \frac{G}{2}\right\},$$

and let $\mathcal{E} := \bigcap_{t=1}^{M-1}\bigcap_{m \notin S_{t-1}^\star} \mathcal{E}_{t,m}$. These events are defined on the deterministic population greedy path before the empirical path is run, so no conditioning on a data-dependent previous-selection event is needed.

On $\mathcal{E}$, the empirical and population paths agree by induction. Suppose $\hat{S}_{t-1} = S_{t-1}^\star$. For any remaining competitor $m \ne m_t^\star$, the population margin and the two strict $G/2$ error bounds give

$$\begin{aligned}&\Delta\widehat{J}_n(m_t^\star \mid S_{t-1}^\star) - \Delta\widehat{J}_n(m \mid S_{t-1}^\star) \\ &\quad> \Delta J(m_t^\star \mid S_{t-1}^\star) - \Delta J(m \mid S_{t-1}^\star) - G \ge 0.\end{aligned}$$

Thus the empirical rule selects $m_t^\star$. Starting from $S_0^\star = \hat{S}_0 = \emptyset$, induction proves equality of the paths through step $M - 1$; the final step has only one candidate.

Step 3: one unconditional union bound. Step 1 and a union bound over the deterministic family of events yield

$$\Pr(\mathcal{E}^c) \le 2 \sum_{t=1}^{M-1} (M-t+1) \exp\left(-\frac{nG^2}{32\sigma^2}\right)$$
$$= \left[M(M+1)-2\right] \exp\left(-\frac{nG^2}{32\sigma^2}\right).$$

Since $\mathcal{E}$ implies recovery of the full ordering, this is the claimed failure-probability bound. Solving it for a target failure probability ϵ proves the sample-complexity statement. ■

Remark. The polynomial prefactor is a conservative consequence of a uniform union bound along the population greedy path. The exponential dependence on nG^2/σ^2 and the logarithmic dependence on M are the substantive parts of the result.

A.3 Experimental Details, Theorems, and Full Tables

This appendix expands the experimental sections of the main paper (Sections 2, 4). It contains the configuration, the formal risk decomposition, the proof of the conditional marginal-utility crossing (Theorem 1), and the full results tables underlying the text figures.

A.3.1 Experimental configuration

Taxonomies (full). The main paper introduces two taxonomies that recur throughout the experiments. **Encoder paradigm** partitions pre-trained vision backbones into five categories: (A) *language-supervised* image–text contrastive models (CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023)); (B) *self-supervised* models trained without labels (DINOv2 (Oquab et al., 2024), MAE (He et al., 2022), Data2Vec (Baevski et al., 2022)); (C) *generative / image-prior* encoders (Stable-Diffusion VAE (Rombach et al., 2022)); (D) *class-supervised* ImageNet-trained models (ConvNeXt (Liu et al., 2022), ResNet-50 (He et al., 2016), ViT-B/16 (Dosovitskiy et al., 2021)); (E) *task-expert* encoders trained for dense vision tasks (SAM (Kirillov et al., 2023)). **Task recognition regime** partitions the five evaluation datasets one-to-one: *Generic Object* (CIFAR-100), *Fine-grained Object* (Pets), *Texture / Material* (DTD), *Geo-spatial / Scene* (Country211), and *Structured Symbol / Digit* (GTSRB).

Reference-path justification. The diagnostic reference path CLIP → DINOv2 → ConvNeXt → SigLIP → MAE → ResNet-50 is fixed a priori from published fusion practice and not optimised on the evaluation data. Its first three positions are the Cambrian-1 (Tong et al., 2024) trio (CLIP, DINOv2, ConvNeXt), reported there as a strong vision-centric combination (Shi et al., 2025; Kar et al., 2024; Zong et al., 2024); the last three add a within-paradigm second representative each (SigLIP for language supervision, MAE for self-supervision, ResNet-50 for class supervision), so positions 5–6 act as redundancy probes against positions 1–3. Self-supervised purely-visual features (DINOv2) are added early to recover what contrastive pre-training misses (Liu et al., 2023).

Hardware. All experiments are conducted on a single NVIDIA RTX 5090 GPU with 32GB VRAM.

Model pool. We use 10 frozen pre-trained encoders spanning all five pre-training paradigms, consistent with Section 4 of the main paper. The language-supervised group contains CLIP ViT-B/16 (Radford et al., 2021) and SigLIP B/16 (Zhai et al., 2023). The self-supervised group contains DINOv2-base (Oquab et al., 2024), ViT-MAE-base (He et al., 2022), and Data2Vec (Baevski et al., 2022). The remaining groups contain the Stable-Diffusion VAE encoder (Rombach et al., 2022) (generative); ConvNeXt-base (Liu et al., 2022), ResNet-50 (He et al., 2016), and ViT-B/16 (Dosovitskiy et al., 2021) (class-supervised); and SAM ViT-B (Kirillov et al., 2023) (task-expert). The generative and task-expert encoders are reintroduced here as adversarial candidates that a competent selector should reject. Feature dimensions range from 256 (SD-VAE) to 2048 (ResNet-50).

Datasets. Both the diagnostic protocol (Section 2, 6-encoder pool) and the results protocol (Section 4, 10-encoder pool) evaluate on the same five datasets, one per recognition regime: CIFAR-100 (100 classes, generic object), Oxford Pets (37 classes, fine-grained object), DTD (47 classes, texture), Country211 (211 classes, geo-spatial scene) and GTSRB (43 classes, structured symbol/digit). Oxford Flowers-102 additionally appears only in the fair 16-shot comparison against single-model CLIP adaptation (Table A.3), following the standard few-shot adaptation benchmark. All datasets are public benchmarks used under their original train/test splits.

Setting B (Concat + MLP). Features from selected encoders are concatenated (no normalization) and fed to a 2-layer MLP: Linear(input_dim $\to 512$) $\to$ ReLU $\to$ Dropout(0.1) $\to$ Linear$(512 \to$ num_classes). Training: Adam optimizer, lr $= 10^{-3}$, weight decay 10^{-4}, batch size 128, 20 epochs.

Selection cost. KAGES needs only the centred encoder kernels and a greedy traversal, with no downstream training. Computing the ten $n \times n$ centred kernels takes $\sim$5s on a single GPU, and the full greedy ordering (three Frobenius inner products per candidate per step, $O(n^2)$ and independent of encoder dimension) completes in under one second. The total selection cost is a few seconds, against hours for exhaustive evaluation of all $2^{10} - 1 = 1023$ subsets.

Reproducibility. Main-table experiments use five random seeds $\{17, 23, 42, 71, 137\}$ and report mean results; aggregate variability is summarised below.

A.3.2 Risk decomposition (Theorem A.3) and proof

For a selected set S, let

$$\bar{F}_S := (F_m)_{m \in S}$$

denote the joint collection of its encoder features. Let $R^\star(S)$ be the Bayes risk given $\bar{F}_S$, and let $R_n(S)$ be the expected population test risk of the predictor returned by the n-sample training procedure. Define the finite-sample excess risk

$$E_n(S) := R_n(S) - R^\star(S) \geq 0.$$

This exact excess term may absorb estimation, approximation, and optimization effects of the specified downstream training procedure.

Theorem A.3 (Risk decomposition)**.** *For a nested growth path $S_0 \subset S_1 \subset \cdots \subset S_K$, the change in finite-sample risk at step k decomposes as*

$$R_n(S_k) - R_n(S_{k-1}) = \Delta E_k - \delta_k^\star,$$

where

$$\Delta E_k := E_n(S_k) - E_n(S_{k-1}), \qquad \delta_k^\star := R^\star(S_{k-1}) - R^\star(S_k) \geq 0.$$

Proof. Subtract $R_n(S_{k-1}) = R^\star(S_{k-1}) + E_n(S_{k-1})$ from $R_n(S_k) = R^\star(S_k) + E_n(S_k)$. The non-negativity of $\delta_k^\star$ follows because $\bar{F}_{S_{k-1}}$ is a subcollection of $\bar{F}_{S_k}$: a Bayes decision rule given $\bar{F}_{S_k}$ can ignore the newly added feature and reproduce any rule based on $\bar{F}_{S_{k-1}}$. The exact increment ΔE_k need not be non-negative. □

A.3.3 Marginal-utility crossing: proof of Theorem 1

This appendix gives the derivation behind Theorem 1 of the main paper. Along a fixed ordering, write

$$g_k := R^\star(S_{k-1}) - R^\star(S_k), \qquad c_k := E_n(S_k) - E_n(S_{k-1}), \qquad a_k := g_k - c_k.$$

Step 1: one-step identity. For every k,

$$R_n(S_k) - R_n(S_{k-1}) = c_k - g_k = -a_k.$$

This follows immediately by subtracting the two risk decompositions in Theorem A.3.

Step 2: log loss makes g_k a conditional mutual information. Under log loss $\ell(q, y) = -\log q(y)$, the Bayes minimizer is the true conditional distribution and

$$R^\star_{\log}(S) = H(Y \mid \bar{F}_S).$$

Since $\bar{F}_{S_k} = (\bar{F}_{S_{k-1}}, F_{m_k})$,

$$g_{k,\log} = I(F_{m_k}; Y \mid \bar{F}_{S_{k-1}}).$$

The chain rule gives

$$I(F_{m_k}; Y \mid \bar{F}_{S_{k-1}}) = \underbrace{I(F_{m_k}; Y)}_{\text{relevance}} - \underbrace{I(F_{m_k}; \bar{F}_{S_{k-1}})}_{\text{redundancy}} + \underbrace{I(F_{m_k}; \bar{F}_{S_{k-1}} \mid Y)}_{\text{class-conditional adjustment}}.$$

Step 3: a submodular greedy path has non-increasing Bayes gains. Let $f(S) := I(\bar{F}_S; Y)$ and suppose f is submodular along the path. If

$$m_k \in \arg\max_{m \notin S_{k-1}} \Delta f(m \mid S_{k-1}),$$

then

$$g_{k+1} = \max_{m \notin S_k} \Delta f(m \mid S_k) \le \max_{m \notin S_k} \Delta f(m \mid S_{k-1}) \le g_k.$$

Thus submodularity plus greedy selection is one sufficient mechanism for the non-increasing-g_k assumption in Theorem 1.

Step 4: Gaussian OLS instance with an increasing pure-noise tail. Let $X \sim \mathcal{N}(0, I_M)$, $Y = \sum_{j=1}^{r} \beta_j X_j + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $F_{m_j}(X) = X_j$, and $S_k = \{m_1, \dots, m_k\}$. For an OLS probe trained on $n > k + 1$ samples,

$$\mathbb{E}\, R_n(S_k) = \nu_k^2 \frac{n-1}{n-k-1}, \qquad \nu_k^2 := \sigma^2 + \sum_{j=k+1}^{r} \beta_j^2.$$

Here $R^\star(S_k) = \nu_k^2$, so

$$E_n(S_k) = \nu_k^2 \frac{k}{n-k-1}.$$

For every pure-noise addition $k > r$, $\nu_{k-1}^2 = \nu_k^2 = \sigma^2$, and hence

$$c_k = E_n(S_k) - E_n(S_{k-1}) = \sigma^2 \frac{n-1}{(n-k)(n-k-1)} > 0.$$

This quantity is strictly increasing with k while $n > k+1$. Thus the Gaussian model gives an explicit increasing-cost tail; it does not assert that c_k is non-decreasing for arbitrary signal-bearing additions. As a concrete full-path instance, take $r = 1$ and $n > K + 1$. Then $g_1 = \beta_1^2$, $c_1 = \sigma^2/(n-2)$, and for every $k \ge 2$, $g_k = 0 < c_k$. Therefore $\beta_1^2 > \sigma^2/(n-2)$ yields a strict single crossing after the first step.

Step 5: proof of the conditional crossing theorem. Under the assumptions of Theorem 1, g_k is non-increasing and c_k is non-decreasing, so $a_k = g_k - c_k$ is non-increasing. Let

$$k^\star := \max\{k : a_k > 0\}$$

and let $\eta := \min_k |a_k| > 0$ denote the no-tie margin. Then $a_k \ge \eta > 0$ for $k \le k^\star$ and $a_k \le -\eta < 0$ for $k > k^\star$; the sign change occurs between $k^\star$ and $k^\star + 1$. By Step 1, $R_n(S_k) - R_n(S_{k-1}) = -a_k$, which is strictly negative up to $k^\star$ and strictly positive afterwards. This proves the asserted order-specific peak-then-decline path. □

A.3.4 Exact CKA margin derivation

We give the full derivation of Proposition 3 in the main paper. Recall the five scalars $A_S := \langle K_S^c, K_Y^c \rangle_F$, $B_S := \langle K_S^c, K_S^c \rangle_F$, $r_m := \langle K_m^c, K_Y^c \rangle_F$, $h_m := \langle K_m^c, K_S^c \rangle_F$, $q_m := \langle K_m^c, K_m^c \rangle_F$, with $C = \|K_Y^c\|_F^2$. This derivation concerns the pure alignment utility U, equivalently J with $\tau = 0$.

One-step CKA after adding m. Adding m to S gives $K^c_{S\cup\{m\}} = K^c_S + K^c_m$, so

$$A_{S\cup\{m\}} = A_S + r_m, \quad B_{S\cup\{m\}} = B_S + 2h_m + q_m, \quad U(S\cup\{m\}) = \frac{A_S + r_m}{\sqrt{(B_S + 2h_m + q_m)\,C}}.$$

Positive-gain condition. Assume $B_S > 0$ and $C > 0$. Since the centred Gram matrices are positive semidefinite, $A_S, r_m, h_m, q_m \geq 0$. Then $U(S\cup\{m\}) > U(S)$ is equivalent to

$$\frac{A_S + r_m}{\sqrt{B_S + 2h_m + q_m}} > \frac{A_S}{\sqrt{B_S}},$$

which (with $A_S \geq 0$, both sides non-negative) is equivalent to

$$(A_S + r_m)^2 B_S > A_S^2(B_S + 2h_m + q_m),$$

i.e., $r_m^2 B_S + 2A_S r_m B_S > A_S^2(2h_m + q_m)$. Treating this as a quadratic in r_m and taking the positive root yields

$$r_m^\star = A_S\left(\sqrt{1 + \frac{2h_m + q_m}{B_S}} - 1\right),$$

which is the exact criterion of Proposition 3.

A.3.5 Diagnostic-protocol scaling tables

Table A.1 reports the full per-dataset accuracy curves that underlie Figure 1 of the main paper.

Table A.1: 10-shot accuracy (%) as encoders are incrementally added along the literature-tracked path (CLIP → DINOv2 → ConvNeXt → SigLIP → MAE → ResNet-50; gated fusion). These are the per-dataset curves underlying Figure 1 and Table 1. **Bold**: peak. ⋆: CLIP alone is optimal.

Dataset	1	2	3	4	5	6	Peak
CIFAR-100	60.29	80.11	**80.37**	80.18	80.20	79.92	3
Pets	85.04	92.86	**93.75**	93.46	93.65	93.49	3
DTD	64.41	72.67	73.52	74.05	74.17	**74.18**	6
Country211⋆	**13.50**	9.85	9.91	9.82	9.71	9.58	1
GTSRB⋆	**61.15**	51.54	52.73	52.34	51.97	52.58	1

A.3.6 Two failure modes (formal characterization)

The risk decomposition (Theorem A.3) shows that step k degrades performance exactly when the excess-risk increment $c_k = \Delta E_k$ exceeds the Bayes-risk reduction $g_k = \delta_k^\star$. Two empirically distinct mechanisms can make this net utility negative.

Type I: Task-conditional redundancy.

Definition A.1 (Task-conditional redundancy)**.** Encoder m_k is *task-conditionally redundant* with respect to S_{k-1} for label Y if

$$I(F_{m_k}; Y \mid \bar{F}_{S_{k-1}}) = 0.$$

Proposition A.1 (Type I removes the Bayes gain)**.** *Under log loss, if m_k is task-conditionally redundant with respect to S_{k-1}, then $g_k = \delta_k^\star = 0$ and*

$$R_n(S_k) - R_n(S_{k-1}) = \Delta E_k.$$

Consequently, the step degrades performance exactly when $\Delta E_k > 0$. The same conclusion holds for any Bayes decision loss whenever Y is conditionally independent of F_{m_k} given $\bar{F}_{S_{k-1}}$.

Proof. Under log loss,

$$g_k = H(Y \mid \bar{F}_{S_{k-1}}) - H(Y \mid \bar{F}_{S_k}) = I(F_{m_k}; Y \mid \bar{F}_{S_{k-1}}) = 0.$$

The one-step identity then gives $R_n(S_k) - R_n(S_{k-1}) = \Delta E_k$. □

TYPE II: TASK–MODEL MISALIGNMENT.

Definition A.2 (Task–model misalignment)**.** For a tolerance $\epsilon \geq 0$, encoder m_k is *ϵ-task-misaligned* relative to S_{k-1} if its conditional Bayes gain satisfies

$$g_k = R^\star(S_{k-1}) - R^\star(S_k) \leq \epsilon.$$

Under log loss this is equivalent to $I(F_{m_k}; Y \mid \bar{F}_{S_{k-1}}) \leq \epsilon$. The candidate may nevertheless be representationally distinct according to a task-agnostic diversity criterion.

Proposition A.2 (Type II decline when excess cost dominates)**.** *If m_k is ϵ-task-misaligned relative to S_{k-1} and $c_k > \epsilon$, then*

$$R_n(S_k) - R_n(S_{k-1}) = c_k - g_k \geq c_k - \epsilon > 0.$$

Proof. The claim follows directly from the one-step identity and the definition $g_k \leq \epsilon$. □

A.3.7 IMAGE RETRIEVAL: FULL RECALL@1 CURVES

With the same KAGES ordering used in the main text, L2-normalized concatenated features, and nearest-neighbor search, we report Recall@1 as a function of k on three datasets (Table A.2).

Table A.2: Image retrieval Recall@1 (%) vs. k. Retrieval gains are task-dependent and saturate at larger k than classification.

Dataset	$k=1$	3	5	8	10	Peak k	Best R@1
DTD	47.4	66.6	70.1	70.5	**74.6**	10	74.6
GTSRB	50.4	56.8	55.7	59.2	**60.2**	10	60.2
Pets	89.8	93.3	93.5	**94.3**	94.0	8	94.3

A.3.8 COMPARISON WITH PUBLISHED SINGLE-MODEL ADAPTERS

Table A.3 compares multi-model fusion with recent single-model CLIP adaptation methods using published 16-shot results for context, without controlling all pretraining and tuning resources.

Table A.3: 16-shot comparison: multi-model fusion vs. single-model CLIP adaptation (%). Adaptation and fusion rows use 16 shots per class; CLIP zero-shot is a separate reference. †: from original papers.

	Method	DTD	Flowers	GTSRB	Avg
Single	CLIP zero-shot†	44.3	67.4	–	–
	CoOp† (Zhou et al., 2022)	68.7	95.7	–	–
	Tip-Adapter-F† (Zhang et al., 2022)	73.7	97.2	–	–
	TaskRes† (Yu et al., 2023)	73.4	97.0	–	–
	CaFo† (Zhang et al., 2023)	74.6	97.5	–	–
Multi	All Models	77.9	99.6	61.6	79.7
	Random	77.7	99.6	60.3	79.2
	KAGES	**78.7**	**99.7**	**72.1**	**83.5**

A.3.9 FULL LEADERBOARD ON PROTOCOL B

The main paper reports KAGES against the standard multi-view defaults (full fusion, random, diversity) and summarises the two stronger subset-selection baselines (DPP, facility location) in the text. Table A.4 gives the complete per-dataset AULC for all six multi-view methods on Protocol B (16-shot, Concat+MLP), means over five seeds. KAGES is best on every dataset and on the average; the coverage-driven DPP and facility-location objectives are the strongest competitors but still trail KAGES, since neither uses the label kernel.

Table A.4: Full per-dataset AULC (%) on Protocol B (16-shot, Concat+MLP), means over 5 seeds $\{17, 23, 42, 71, 137\}$. **Bold**: best per column; underlined: second-best.

Method	C100	Pets	DTD	C211	GTSRB	Avg
Fuse-All	78.54	94.06	75.35	10.16	57.18	63.06
Random	77.06	92.97	74.87	11.10	59.16	63.03
Diversity	80.61	90.91	74.38	11.27	59.12	63.26
DPP	83.62	93.21	72.48	11.87	64.88	65.21
Facility location	82.82	93.31	77.64	11.34	65.53	66.13
KAGES (ours)	**83.84**	**94.70**	**77.82**	**12.16**	**66.12**	**66.93**

A.3.10 Normalization ablation

The main paper uses the raw additive variant $K_S^c = \sum_{s \in S} K_s^c$. A per-encoder Frobenius-normalised variant $\widetilde{K}_s^c := K_s^c / (\|K_s^c\|_F + \varepsilon)$ gives provable per-encoder scale invariance but yields slightly lower empirical AULC on our pool ($\Delta\mathrm{AULC} \approx -0.4$ pp on Protocol B). We report it here as an ablation and use the raw additive variant in the main results.

A.3.11 Inter-seed stability of AULC

Across the five seeds $\{17, 23, 42, 71, 137\}$, the inter-seed standard deviation of AULC is reported to be small: it is at most 0.3 pp for KAGES on every dataset and below 0.6 pp for all baselines, although these aggregate bounds alone do not establish per-seed rankings or paired significance.

A.4 Cross-Modal Validation on Frozen LLMs

To test whether KAGES transfers beyond vision, we run a parallel experiment on **6 frozen base LLMs**: Qwen2.5-7B (Yang et al., 2024), Mistral-7B-v0.1, Llama-3-8B (Dubey et al., 2024), DeepSeek-7B, Gemma-7B, and Yi-1.5-9B. For each model we extract the last-layer, last-token hidden state on MMLU-500 and ARC-Challenge-500 (both 4-way MCQA, treated as classification), then apply the exact same pipeline as our main experiments: concatenate features of the first k selected models, train a linear probe on an 80/20 split, and report accuracy (Table A.5).

Table A.5: Linear-probe accuracy (%) across k on frozen LLMs. Peak-then-decline reproduces on both tasks. KAGES peaks earlier and higher on MMLU.

Dataset	Method	$k=1$	2	3	4	5	6	Peak
MMLU-500	**KAGES**	67.0	**75.0**	69.0	72.0	71.0	67.0	**75.0**
	diversity-greedy	67.0	66.0	70.0	69.0	71.0	68.0	71.0
	fixed order	67.0	68.0	72.0	70.0	69.0	67.0	72.0
ARC-Ch-500	**KAGES**	88.0	86.0	89.0	**91.0**	88.0	87.0	**91.0**
	diversity-greedy	88.0	86.0	89.0	90.0	89.0	87.0	90.0
	fixed order	85.0	88.0	89.0	85.0	91.0	89.0	91.0

Findings. (i) The peak-then-decline curve reproduces: MMLU peaks at $k = 2$ for KAGES (75%) and at $k = 3$ to 5 for the baselines, while ARC peaks at $k = 4$ to 5. No configuration benefits from all 6 models. (ii) KAGES's alignment score beats diversity-greedy by up to 9pp at $k = 2$ on MMLU, consistent with our paper observation that pure diversity is necessary but not sufficient. (iii) The CKA matrix supports the Platonic Representation Hypothesis (Huh et al., 2024) across modalities: Mistral and Llama, both English-dominant base models, cluster tightly, while Gemma has comparatively low similarity to the other models.

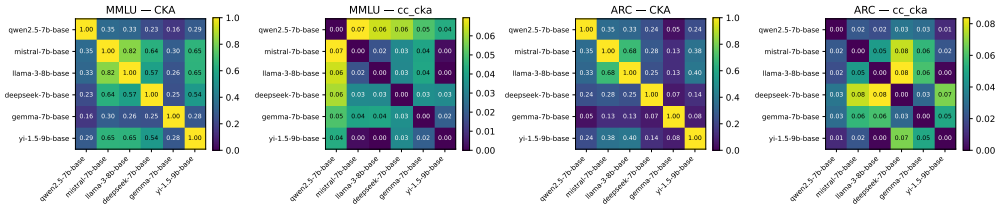

Figure A.1: **CKA matrices across 6 LLMs.** Mistral-Llama CKA reaches 0.82 on MMLU (both English-heavy pretraining), while Gemma has relatively low pairwise CKA with the other models. These similarities do not isolate the effects of architecture, training corpus, or objective.

A.5 Extended Related Work

Foundation-model fusion and transferability scoring. Classical multi-view learning (Xu et al., 2013; Yan et al., 2021; Trosten et al., 2021) fuses a small fixed set of feature views and offers no recipe for selecting views out of a heterogeneous pool. Multi-encoder MLLMs scale this to multiple frozen vision backbones: Eagle (Shi et al., 2025), Cambrian-1 (Tong et al., 2024), BRAVE (Kar et al., 2024), and MoVA (Zong et al., 2024) study encoder combinations or routing within relatively small, predefined expert pools. Training-free transferability scores (LogME (You et al., 2021; 2022), H-Score (Bao et al., 2019), GBC (Pandy et al., 2022), LEEP (Nguyen et al., 2020), NCE (Tran et al., 2019), TransRate (Huang et al., 2022), EMMS (Meng et al., 2023), PACTran (Ding et al., 2022), OTCE (Tan et al., 2021; 2024), DISCO (Zhang et al., 2025)) score one encoder and do not model its interaction with a selected pool. Pool-level approaches are sparse: OSBORN (Vimal et al., 2023) gives a fixed-budget guarantee, while model-zoo studies (Dong et al., 2022; Chen et al., 2023) select at the recipe level. KAGES instead constructs a complete task-aware ordering over a heterogeneous pool.

Kernel and information selection. Classical feature selection (mRMR (Peng et al., 2005), JMI (Yang & Moody, 1999), Brown et al. (Brown et al., 2012)) and similarity tools (CKA (Kornblith et al., 2019), HSIC (Gretton et al., 2005), Platonic Hypothesis (Huh et al., 2024)) compare representations or select scalar features but do not optimise prefix-path quality. Greedy kernel selectors such as HSIC-mRMR, HSIC-Lasso (Song et al., 2012; Yamada et al., 2014), and Kernel-JMI (Brown et al., 2012) select scalar features through pairwise scores whose cost grows with feature dimension. KAGES treats each encoder representation as one atomic unit, evaluates candidates directly at the kernel level, and adapts its relevance-redundancy balance to the current selected set.

A.6 Future Work: Online and Sequential Extensions

The present framework is offline: the labelled task is fixed at n samples, the encoder pool is finite, and the selection is computed in a single pass. Two extensions are natural future directions.

Sequential / streaming labels. If labelled samples arrive sequentially, the ordering $\hat{\pi}$ and the optional score-selected size $\hat{k}_J$ can be recomputed as the empirical score path tightens. The sequential-analysis literature (Tartakovsky et al., 2014) provides change-point detection tools that could be adapted to detect significant shifts in the selection. Active best-arm identification (Kaufmann et al., 2016) provides sample-complexity templates for adaptive selection.

Online ordering and score-based size selection. The ordering and the optional score-selected size can be recomputed as new samples arrive. Under its stated assumptions, Theorem A.2 implies stabilization of the *ordering* after $O((B^2 + \sigma^2[\log M + \log(1/\epsilon)])/G^2)$ samples with failure probability at most ϵ; it does not by itself prove stabilization or accuracy optimality of $\hat{k}_J$. An online variant could expose intermediate orderings and score paths with confidence intervals. The connection to bandit best-arm identification (Even-Dar et al., 2002; 2006; Kaufmann et al., 2016) is natural and we leave the formal analysis to a follow-up.